Nonparametric Regression under Alternative Data Environments

Abdoul G. Sam and Alan P. Ker *

Abstract

This paper proposes a nonparametric regression estimator which can accommodate two empirically relevant data environments. The first data environment assumes that at least one of the explanatory variables is discrete. In such an environment a "cell" approach which consists of partitioning the data and estimating a separate regression for each discrete cell has usually been employed. The second data environment assumes that one needs to estimate a set of regression functions that belong to different experimental units. In both environments the proposed estimator attempts to reduce estimation error by incorporating extraneous data from the other experimental units or cells when estimating the regression function for a given individual experimental unit or cell. Consistency and asymptotic normality of the proposed estimator are established. Its computational simplicity and simulation results demonstrate a strong potential in empirical applications.

Keywords: Extraneous information, bias reduction, correction factor, data environments.

*The authors are Ph.D. candidate and assistant professor, respectively, in the Departments of Economics and Agricultural and Resource Economics, University of Arizona. Corresponding author: Alan Ker, Associate Professor and Head, Department of Agricultural and Resource Economics, University of Arizona, Tucson, AZ, 85721-0023. Tel: (520) 621 6242. Fax: (520) 621 6250. E-mail:aker@ag.arizona.edu.

1 Introduction

Let (X_{ij}, Y_{ij}), $i = 1, ..., n_j, j = 1, ..., Q$ be a sample of R^{p+1} valued random vectors where Y_{ij} represents a response variable and X_{ij} is a p-dimensional vector of explanatory variables. In many empirical situations it is necessary to estimate a set of regression curves, say one for each experimental unit of interest, which can be arranged as

$$Y_{ij} = m_j(X_{ij}) + \epsilon_{ij} \tag{1}$$

where j denotes the j^{th} experimental unit, and ϵ_{ij} is a zero-mean and finite-variance error process. In this manuscript, we concern ourselves with the estimation of the conditional mean $E(Y|X = x)$. Kernel regression estimators based solely on individual samples such as the Nadaraya-Watson and the local linear kernel estimators have become widespread because they circumvent the risk of functional misspecification inherent to their parametric counterparts and provide consistent estimates under mild regularity conditions.

The standard Nadaraya-Watson estimator of the conditional mean $m_j(x)$ is given by

$$\tilde{m_j}(x) = \frac{\sum_{i=1}^{n_j} Y_{ij} K_{h_j}(X_{ij} - x)}{\sum_{i=1}^{n_j} K_{h_j}(X_{ij} - x)} \tag{2}$$

where h_j is the smoothing parameter and $K_{h_j}(u) = \frac{1}{h_j} K(\frac{u}{h_j})$ with $K(u)$ being the kernel function. Denoting $\mu_2 = \int z^2 K(z) dz$ and $\mathrm{R}(K) = \int K^2(z) dz$, the standard properties of the Nadaraya-Watson estimator are

$$E[\tilde{m_j}(x) - m_j(x)] = \frac{1}{2}\mu_2 h_j^2 \{m_j''(x) + 2m_j'(x)\frac{f_j'(x)}{f_j(x)}\} + o(h_j^2), \tag{3}$$

$$\mathrm{V}ar[\tilde{m_j}(x)] = \frac{\sigma^2 R(K)}{(nh_j) f_j(x)} + O(h_j/n_j) \tag{4}$$

where $f_j(x)$ is the marginal density function of X_{ij} evaluated at support point x. Since the bias is $O(h_j^2)$ and $h_j = h_j(n_j)$ goes to 0 as n_j goes to ∞, it follows that the Nadaraya-Watson estimator is consistent. However, a drawback is its finite sample bias which can be quite large. Several papers have proposed estimators which reduce the bias (Hardle and Browman, 1988; Hjort and Glad, 1995; Glad, 1998, among others) or eliminate it (Racine, 2001). One such bias-correction estimator, Hjort and Glad (1995), is of particular interest because of the ease with which it can be implemented. Hjort and Glad (1995) propose a semiparametric estimator which combines a parametrically estimated pilot with a nonparametrically estimated correction factor. The parametric pilot can be thought of as a prior for the shape of $m_j(x)$ whereas the correction factor adjusts the pilot if it does not satisfactorily capture the shape of $m_j(x)$. Consequently, the estimator behaves like the parametric start if the parametric assumption is correct, while resembling the nonparametric estimator otherwise.

The estimator we propose in this manuscript is in the same realm as Hjort and Glad's (1995); however, we consider alternative data environments where we have data from possibly similar regression functions. If those unknown functions are identical, the optimal estimator would pool the data and estimate one regression curve. If, however, those unknown functions are sufficiently similar, using the pooled estimator as a pilot in Hjort and Glad's framework would yield efficiency gains relative to the Nadaraya-Watson estimator. The use of extraneous data in the form a nonparametric pooled start represents the key conceptual difference between our proposed estimator and the estimator of Hjort and Glad (1995).

Two empirically relevant data environments are considered. The first data environment assumes that at least one of the explanatory variables is discrete. While this situation is easily accommodated in a parametric framework, the continuity assumptions required for nonparametric regression are violated. As a result, a separate nonparametric regression estimation is required for each discrete value. For example, if one of the explanatory variables is discrete and may take values $\{0, 1, 2, 3\}$, the sample must be partitioned according to the four discrete values into four cells where a separate regression function is undertaken for each. Recently, Racine and Li (2004) developed a nonparametric estimator that smoothes across the discrete values, thereby reducing variance at a cost of increased bias. Conversely, our proposed estimator attempts to reduce bias by utilizing the entire data set. The second data environment assumes that one needs to estimate a set of regression curves rather than a single regression curve. Empirically, this situation arises often and led Altman and Casella (1995) to develop a Stein-type Bayesian nonparametric estimator that uses empirical Bayes techniques pointwise across the function space to reduce estimation error. This latter data environment can be viewed as a generalization of the former with each of the discrete cells representing an experimental unit.

The remainder of this manuscript is organized as follows. In the second section we introduce the proposed estimator and investigate its asymptotic properties. The third section presents our simulation results. The final section summarizes our findings.

2 A Nonparametric Estimator with a Pooled Start

Underlying the proposed estimator is that there exists a prior belief that the conditional means are similar in shape. If the curves were identical, that is, if $m_1(x) = m_2(x) = ...m_Q(x) = m(x)$, we would simply pool the data and estimate one common curve. Conversely, if the conditional means were dissimilar, the pooled estimator is inappropriate. A primary strength of the proposed estimator is that the form or extent of similarity among the curves is not required; in most empirical applications the form or extent of similarity is unknown. We have adapted the Hjort and Glad estimator to the context of model (1) by combining pooled and individual nonparametric estimators. As a result, the proposed estimator, which we denote the nonparametric estimator with a pooled start (NEPS), resembles the pooled estimate if the curves are identical or similar and the individual (Nadaraya-Watson) estimate if the curves are dissimilar. The NEPS estimator of conditional mean $m_j(x)$ is

$$\hat{m}_j(x) = \hat{m}_p(x)\hat{r}_j(x) = \frac{\sum_{i=1}^{n_j} Y_{ij}[\frac{\hat{m}_p(x)}{\hat{m}_p(X_{ij})}]K_{h_j}(X_{ij} - x)}{\sum_{i=1}^{n_j} K_{h_j}(X_{ij} - x)}. \tag{5}$$

The estimator is implemented in two steps. The first step pools the data from all experimental units to estimate a single curve denoted $\hat{m}_p(x)$. This step introduces extraneous information from the pooled dataset that is potentially relevant to the estimation of the conditional mean of interest. The second step consists of multiplying the pooled estimate by a nonparametrically estimated correction factor $\hat{r}_j(x)$ to account for individual effects. The NEPS estimator is designed to outperform the standard Nadaraya-Watson estimator when the hypothesis of similarity is tenable, but also produce reliable estimates when the curves are dissimilar.

2.1 Asymptotic Properties of the NEPS Estimator

In deriving the asymptotic properties of the NEPS estimator, we require the following assumptions:
A1. The X_{ij}s are i.i.d. and independent of the error process ϵ_{ij}, which is also i.i.d.

A2. The density function $f_j(x)$ and the conditional mean $m_j(x) \in \mathcal{C}^2(\Theta)$ with finite second derivatives and $f_j(x) \neq 0$ in Θ, the neighborhood of point x.
A3. The density function $g(x)$ and the conditional mean $m_p(x)$ of the pooled data $\in \mathcal{C}^2(\Theta)$ with finite second derivatives and $g(x) \neq 0$ in Θ, the neighborhood of point x. The density function of the pooled data $g(x)$ is a mixture density generated by the Q individual density functions, that is, $g(x) = \sum_{j=1}^{Q} w_j f_j(x)$ where w_j are mixing weights.
A4. The kernel function $K(z)$ is bounded, real-valued, with the following characteristics: (i) $\int K(z)dz = 1$, (ii) $K(z)$ is symmetric about 0, (iii) $\int z^2 K(z)dz < \infty$, (iv) $|z|K(|z|) \to 0$ as $|z| \to \infty$, (v) $\int K^2(z)dz \leq \infty$.
A5. $h_j \to 0$ and $n_j h_j \to \infty \ \forall \ j = 1, ..., Q$.
A6. $E|\epsilon_i|^{2+\delta}$, $\int |K(\omega)|^{2+\delta}$, and $\int |\frac{m_p(x)}{m_p(X_{ij})}|^{2+\delta}$ are finite for some $\delta > 0$.
A7. We assume that $h_p \to 0$ and $n_j h_p \to \infty \ \forall \ j = 1, ..., Q$ where h_p is the smoothing parameter for the pooled estimator.

Theorem

1. Under assumptions A1-A5, we have

$$E[\hat{m}_j(x) - m_j(x)] = \frac{1}{2}\mu_2 h_j^2 \{r_j''(x) + 2r_j'(x)\frac{f_j'(x)}{f_j(x)}\} m_p(x) + o(h_j^2) \tag{6}$$

$$\mathrm{V}ar[\hat{m}_j(x)] = \frac{\sigma^2 R(K)}{(nh_j)f_j(x)} + O(h_j/n_j + (Nh_p)^{-1}). \tag{7}$$

2. Under assumptions A1-A7, $\hat{m}_j(x)$ has a limiting normal distribution

$$\sqrt{n_j h_j}(\hat{m}_j(x) - m_j(x)) \to N(B(h_j), \Sigma_j) \tag{8}$$

where $B(h_j) = \frac{1}{2}\mu_2 h_j^2 [m_p(x) r_j''(x) + 2m_p(x) r_j' \frac{f_j'(x)}{f_j(x)}]$ and $\Sigma_j = \frac{\sigma^2}{f_j(x)} R(K)$.

Proof: See appendix.

Equations 4 and 7 show that the variances of the Nadaraya-Watson estimator and the NEPS estimator are essentially the same; the two expressions differ by $O(\frac{1}{n_1 h_p + n_2 h_p + ... + n_Q h_p})$, which is asymptotically negligible by A7. The bias of the NEPS estimator is not a function of the slope and curvature of the true regression function as it is for the Nadaraya-Watson estimator (see equation 3). Rather, the bias is a function of the slope and second derivative of the correction factor $r_j(x)$. If the nonparametric pilot m_p coincides with or is proportional to the true function m_j, then $r_j(x)$ will be a straight line and $r_j' = r_j'' = 0$. This implies that the leading terms of the bias will vanish. Similarly if $m_p(x)$ and $m(x)$ are sufficiently similar, the correction factor will be less variable than the individual conditional mean, hence leading to bias reduction. Interestingly, the pooled start does not have to be a good approximation of $m_j(x)$ for the NEPS estimator to remain competitive to the Nadaraya-Watson estimator in moderate samples.

2.2 Computational Tips

The ratio $\frac{\hat{m}_p(x)}{\hat{m}_p(X_{ij})}$ can be highly influential in regions when X_{ij} is far from x. Also, it is possible that $\hat{m}_p(X_{ij})$ and $\hat{m}_p(x)$ have different signs. Following Glad (1998), we suggest substituting the ratio

$\frac{\hat{m}_p(x)}{\hat{m}_p(X_{ij})}$ by

$$\left| \frac{\hat{m}_p(x)}{\hat{m}_p(X_{ij})} \right|_{\frac{1}{10}}^{10},$$

that is, truncating values below $\frac{1}{10}$ and above 10 to make the estimator robust to these local effects. Additionally, when the number of curves Q is large, selecting the "optimal" extraneous data to be included in the nonparametric pilot is not trivial. This problem is analogous to the choice of instruments in instrumental variable estimation when the number of instruments is large or the choice of the functional form of the parametric pilot in Hjort and Glad (1995). A cross-validation procedure to select the extraneous data to be included in the pooled start can be used. The cross-validation procedure consists of alternating the pooled start from the set formed by the Q curves, for a total of 2^Q possible pooled guides, and then choosing the one whose loss function is the lowest.

3 Finite Sample Simulations

In this section we conduct simulations to investigate the empirical applicability of the NEPS estimator compared to the Nadaraya-Watson and other related estimators. Prior to the simulation results we provide a terse review of two related estimators.

3.1 The Racine and Li Estimator

The objective of the Racine and Li estimator is to nonparametrically estimate regression functions with discrete independent variables without having to partition the data. Suppose we have data on one experimental unit: Y_i a scalar response variable, X_i^c a vector of continuous variables, and X_i^d an r-dimensional vector of discrete regressors. The Racine and Li estimator smoothes the continuous variables by a c-variate kernel while the discrete variables are smoothed as follows

$$S(X_{it}^d, x_t^d, \lambda) = \begin{cases} 1 & \text{if } X_{it}^d = x_t^d \\ \lambda & \text{otherwise} \end{cases}, 0 \leq \lambda \leq 1 \tag{9}$$

where X_{it}^d is the t^{th} component of the vector X_i^d. The Racine and Li estimator is

$$\tilde{m}^{RL}(x^c, x^d) = \frac{\sum_{i=1}^n Y_i W_{h,\lambda}(X_i^c, x^c, X_i^d, x^d)}{\sum_{i=1}^n W_{h,\lambda}(X_i^c, x^c, X_i^d, x^d)} \tag{10}$$

where $W_{h,\lambda}(X_i^c, x^c, X_i^d, x^d) = K_h(X_i^c - x^c) \prod_{l=1}^r S(X_{it}^d, x_t^d, \lambda)$.

In a context of multiple curve estimation as laid out in equation (1), the "discrete" smoother $S(., \lambda)$ controls the inclusion of extraneous information by assigning a weight of 1 to observations belonging to the experimental unit of interest and a weight of λ to observations from the remaining experimental units. The boundedness of λ within the unit interval allows the Racine and Li estimator to nest both the pooled (λ=1) and Nadaraya-Watson (λ=0) estimators.

3.2 The Altman and Casella Estimator

The Altman and Casella model assumes a fixed and balanced design for the predictor variable so that (1) can be rewritten as $Y_{ij} = m_j(X_i) + \epsilon_{ij}$ with $X_i = i/n$. It is also assumed that each curve can be written as $m_j(X_i) = m(X_i) + \eta_j(X_i)$; that is, the curve for experimental unit j at design point X_i is the population mean curve plus a term which captures the deviation from the population mean curve. Underlying this last assumption is the fact that the curves are all sampled from the

same population. Denote $\tilde{m}_j$ the nonparametric estimate of m_j. Given that $\tilde{m}_j$ is biased, it can be expressed as $\tilde{m}_j = \phi_j + \upsilon_j$ where υ_j is an error term such that $E[\upsilon_j] = 0$ and $Var[\upsilon_j] = \alpha^2/n$. Altman and Casella form a hierarchical model ($\tilde{m}_j|\phi_j$ is normally distributed, and ϕ_j and m_j are jointly normally distributed) and derive the posterior mean of m_j as

$$\tilde{m}_j(x) = \bar{m}(x) + \alpha(x)[\tilde{m}_j(x) - \phi(x)]. \tag{11}$$

In practice, the hyperparameters are replaced by sample estimates, which leads to the Altman and Casella estimator for experimental unit j

$$\tilde{m}_j^{AC}(x) = \bar{y}_x + \tilde{\alpha}(x)[\tilde{m}_j(x) - \overline{\tilde{m}}(x)] \tag{12}$$

where $\bar{y}_x = \frac{1}{Q}\sum_{j=1}^{Q} y_{xj}$ is the cross-individual sample mean of the data at design point x, $\tilde{\alpha}(x) = \frac{\tilde{\sigma}_{y(x)m(x)}}{\tilde{\sigma}^2_{m(x)}}$ is the ratio of the covariance between the data and the nonparametric estimates and the variance of the nonparametric estimates, and $\overline{\tilde{m}}(x) = \frac{1}{Q}\sum_{j=1}^{Q}\tilde{m}_j(x)$. The reader is directed to Altman and Casella (1995) for a complete derivation of their model. Note that this estimator uses the data from the other experimental units in the population in the regression of the curve of interest through $\overline{\hat{m}}(x)$ and $\bar{y}_x$. If the individual curves are similar, then $[\hat{m}_i(t) - \overline{\hat{m}}(t)]$ goes to zero and the final estimates behave like $\bar{y}_x$ which is unbiased for the population mean curve. Altman and Casella note that their estimator performs better when the number of experimental units is sufficiently large so that $\bar{y}_x$ provides a good approximation to the population mean.

3.3 Simulation Design

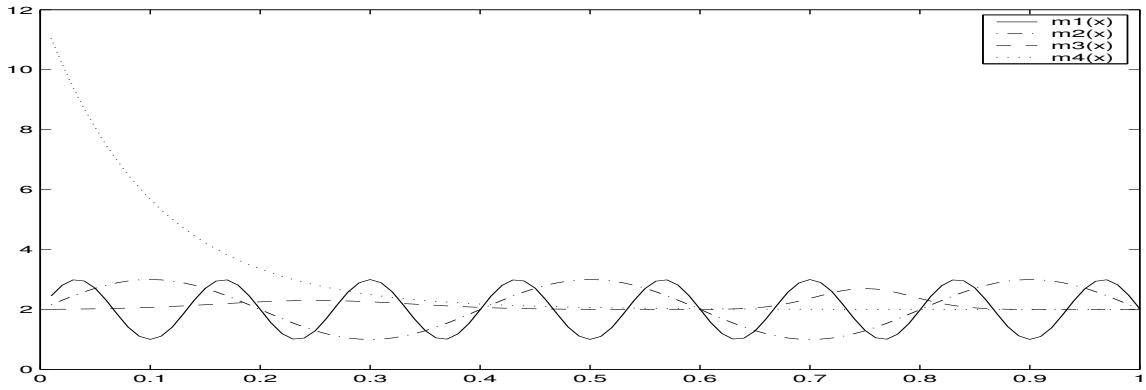
In the first experiment we consider a random design regression where the explanatory variable is uniformly distributed on the [0,1] interval. The second experiment forces the explanatory variable to be equi-spaced on the [0,1] interval as required by the Altman and Casella estimator. For each experiment two scenarios are investigated. In the first scenario, which we denote the "case of identical curves," four identical curves were generated: $m_1(x) = m_2(x) = m_3(x) = m_4(x) = sin(5\pi x)$. Individual-specific errors differentiate the data across experimental units. This is the ideal case for the NEPS estimator. In the second scenario, which we refer to as the "case of dissimilar curves," four very dissimilar curves were generated (see figure 1). The four curves are

$$m_1(x) = sin(15\pi x); \tag{13}$$
$$m_2(x) = sin(5\pi x); \tag{14}$$
$$m_3(x) = .3e^{(-64(x-.25)^2)} + .7e^{(-256(x-.75)^2)}; \text{ and} \tag{15}$$
$$m_4(x) = 10e^{-10x}. \tag{16}$$

Unlike in density estimation where the Marron and Wand densities (1992) are commonly used to study the finite sample performance of density estimators, there are no standard test functions in the regerssion case. However, the curves we use here have also been employed in similar simulations (Hurvitch and Simonoff, 1998; Ruppert, Sheather, and Wand, 1995; Herrmann,1997). The choice of these two extreme scenarios is motivated by the fact that in empirical settings it is impossible to know if the conditional means are similar, identical, or dissimilar. Throughout the simulations, a Gaussian kernel is used and the bandwidth is the one that minimizes the integrated squared error

$$ISE[\hat{m}_j(x)] = \int [\hat{m}_j(x) - m_j(x)]^2 dx. \tag{17}$$

Figure 1: Graph of the four conditional means



3.4 Results

Tables 1 and 2 report the average mean integrated squared error (MISE) of the four curve estimates respectively for the random and fixed designs for samples sizes 50, 100, and 500 with 500 simulations. The average mean integrated squared bias (MIB^2) is also reported as the NEPS estimator is designed to reduce bias. Table 1 presents the results of the first simulation experiment where a random design regression is used. NW denotes the Nadaraya-Watson estimator, R&Li is the Racine and Li estimator, and NEPS is our proposed estimator. For the "case of similar curves" the NEPS estimator significantly outperformed the NW estimator in all sample sizes. The superior performance of the NEPS estimator is attributable to a lower bias as seen in table 1, confirming the derived theory. The R&Li estimator also outperformed the NW estimator but not to the extent of the NEPS. Interestingly, the NEPS estimator also has a lower MISE than the NW estimator for the sample sizes of 50 and 100 in the "case of dissimilar curves." We would expect that as the sample size increases, the NW estimator will perform better than the NEPS estimator when the curves are not identical. An intuitive explanation of this somewhat surprisingly good performance is that the NW estimator is a special case of the NEPS estimator with $m_p(x)$ being equal to a constant $\forall\, x$. However, a "flat start" is quite conservative for most curves, including those curves we consider in these simulations.

Table 1: Average error of the four curves: random design.

Case of similar curves

n	NW		R&Li		NEPS	
	MISE	MIB^2	MISE	MIB^2	MISE	MIB^2
50	10.644	5.6111	3.8884	2.6365	2.6987	0.3431
100	5.7383	3.5223	2.3816	1.7222	1.4638	0.1710
500	1.9022	1.3000	0.5883	.3064	0.3915	0.0553

Case of dissimilar curves

n	NW		R&Li		NEPS	
	MISE	MIB^2	MISE	MIB^2	MISE	MIB^2
50	18.5100	14.3130	20.3510	16.2460	18.1560	12.7090
100	14.8060	12.3190	15.5110	12.9280	14.5970	11.0740
500	12.2757	11.3895	15.1869	14.8708	13.4409	12.2039

Therefore $\hat{m}_p(x)$ need not be a great approximation of the conditional mean of interest for the NEPS estimator to perform well. This result was also found by Hjort and Glad (1995) and Glad (1998) and represents a strength of their idea. Formally, if $m_p(x)$ is such that $|r_j''(x)m_p(x)+2r_j'(x)m_p(x)\frac{f_j'(x)}{f_j(x)}| < |m_j''(x) + 2m_j'(x)\frac{f_j'(x)}{f_j(x)}|$, the NEPS estimator will have a smaller asymptotic mean squared error than the NW estimator as the variances are essentially the same. The R&Li estimator remained competitive because of its ability to revert to the NW estimator by having $\hat{\lambda}_j \to 0$ when the curves are dissimilar.

Table 2 reports the results of the second experiment where a fixed design is used. A&C denotes Altman and Casella's nonparametric empirical Bayes estimator. The NEPS estimator outperformed the NW estimator and the A&C estimator when the conditional means are identical. As in the random design case, the NEPS remained competitive to the NW estimator for the samples sizes of 50 and 100 even when the similarity assumption is inappropriate. The performance of the A&C estimator is somewhat disappointing, which could be explained by the small number of experimental units ($Q = 4$) considered in our simulations. Altman and Casella (1995) noted that Q needs to be large for their estimator to perform well relative to the NW estimator.

Table 2: Average error of the four curves: fixed design.

Case of similar curves

n	NW		A&C		NEPS	
	MISE	MIB^2	MISE	MIB^2	MISE	MIB^2
50	5.6786	1.5812	7.8547	0.0152	2.7299	0.3296
100	3.1335	0.6486	7.4178	0.0149	1.5254	0.1980
500	0.8969	0.2003	6.6718	0.2652	0.6154	0.0273

Case of dissimilar curves

n	NW		A&C		NEPS	
	MISE	MIB^2	MISE	MIB^2	MISE	MIB^2
50	12.4690	8.9958	18.3277	9.1848	11.9579	7.3078
100	4.4908	2.0037	11.1970	3.1989	5.1113	1.7449
500	1.1113	0.2649	11.1993	4.6252	2.7831	0.3238

4 Summary

In this paper, we have proposed a computationally simple nonparametric regression method which admits two empirically relevant data environments. The method was designed to achieve bias reduction by incorporating extraneous information from curves which are thought to be similar to the curve of interest. Consistent with the derived theory, the simulation results indicate that the NEPS estimator has a strong practical potential in small to moderate samples. It outperformed the NW estimator when the curves were identical and did not lose much efficiency when the curves were very dissimilar. The proposed estimator also performed admirably against the related estimators of Racine and Li and Altman and Casella.

References

[1] Altman, N.S., and G. Casella, "Nonparametric Empirical Bayes Growth Curve Analysis," *Journal of the American Statistical Association* 90 (1995), 508-514.

[2] Glad, I., "Parametrically Guided Nonparametric Regression," *Scandinavian Journal of Statistics* 25 (1998), 649-668.

[3] Hardle, W., and S. Browman, "Bootstrapping in Nonparametric Regression: Local Adaptive Smoothing and Confidence Bounds," *Journal of the American Statistical Association* 83 (1988), 102-110.

[4] Herrmann, E., "Local Bandwidth Choice in Kernel Regression Estimation," *Journal of Computational and Graphical Statistics* 6 (1997), 35-54.

[5] Hjort, N.L., and I. Glad, "Nonparametric Density with a Parametric Start," *The Annals of Statistics* 23 (1995), 882-904.

[6] Hurvitch, C., and J.S. Simonoff, "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion," *Journal of Royal Statistical Society Series B* 60 (1998), 271-293.

[7] Marron, J.S., and M.P. Wand, "Exact Mean Integrated Squared Error," *The Annals of Statistics* 20 (1992), 712-736.

[8] Racine, J.S., and Qi Li, "Nonparametric Regression with Both Categorical and Continuous Data," *Journal of Econometrics* 119 (2004), 99-130.

[9] Racine, J. S., "Bias-Corrected Kernel Regression," *Journal of Quantitative Economics* 17 (2001), 25-42.

[10] Ruppert, D., S.J. Sheather, and M.P. Wand, "An Effective Bandwidth Selector for Local Least Squares Regression," *Journal of the American Statistical Association* 90 (1995), 1257-1270.

Appendix. Proof of the Theorem

In what follows we will drop the subscript j for simplicity.

1. Under assumptions A1-A5, we have

$$E[\hat{m}(x) - m(x)] = \frac{1}{2}\mu_2 h^2 \{r''(x) + 2r'(x)\frac{f'(x)}{f(x)}\} m_p(x) + o(h^2) \tag{18}$$

$$\mathrm{V}ar[\hat{m}(x)] = \frac{\sigma^2 R(K)}{(nh)f(x)} + O(h/n + (Nh_p)^{-1}). \tag{19}$$

Proof. $\hat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\left(\frac{Y_i}{\hat{f}(x)}\right)\left(\frac{\hat{m}_p(x)}{\hat{m}_p(X_i)}\right)$. A Taylor series expansion of $\frac{\hat{m}_p(x)}{\hat{m}_p(X_i)}$ around

$\frac{m_p(x)}{m_p(X_i)}$ yields

$$\begin{aligned}\hat{m}(x) &\simeq \frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)\frac{Y_i}{\hat{f}(x)}\left(\frac{m_p(x)}{m_p(X_i)}+\frac{\hat{m}_p(x)-m_p(x)}{m_p(X_i)}\right.\\ &\quad\left.-\frac{m_p(x)}{m_p(X_i)}\frac{\hat{m}_p(X_i)-m_p(X_i)}{m_p(X_i)}\right).\end{aligned}$$

The expressions $\frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)\frac{\epsilon_i}{\hat{f}(x)}\frac{\hat{m}_p(x)-m_p(x)}{m_p(X_i)}$ and $\frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)\frac{\epsilon_i}{\hat{f}(x)}\frac{m_p(x)}{m_p(X_i)}\frac{\hat{m}_p(X_i)-m_p(X_i)}{m_p(X_i)}$ are of order $o_p(h_p^2)$; hence,

$$\begin{aligned}\hat{m}(x) &= \frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)\frac{Y_i}{\hat{f}(x)}\frac{m_p(x)}{m_p(X_i)}+\frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)\frac{m(X_i)}{\hat{f}(x)}\frac{\hat{m}_p(x)-m_p(x)}{m_p(X_i)}\\ &\quad-\frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)\frac{m(X_i)}{\hat{f}(x)}\frac{m_p(x)}{m_p(X_i)}\left(\frac{\hat{m}_p(X_i)-m_p(X_i)}{m_p(X_i)}\right)+o_p(h_p^2)\\ \hat{m}(x)-m(x) &= \frac{m_p(x)}{n\hat{f}(x)}\sum_{i=1}^{n} K_h(X_i-x)(r(X_i)+\epsilon_i^*-r(x))+\frac{1}{n\hat{f}(x)}\sum_{i=1}^{n} K_h(X_i-x)r(X_i)(\hat{m}_p(x)-m_p(x))\\ &\quad-\frac{1}{n\hat{f}(x)}\sum_{i=1}^{n} K_h(X_i-x)\frac{m_p(x)}{m_p(X_i)}r(X_i)(\hat{m}_p(X_i)-m_p(X_i))+o_p(h_p^2)\\ &= \frac{A_n}{\hat{f}(x)}+\frac{B_n}{\hat{f}(x)}+o_p(h_p^2)\end{aligned}$$

where $\epsilon_i^*=\frac{\epsilon_i}{m_p(X_i)}$, $A_n=\frac{m_p(x)}{n}\sum_{i=1}^{n} K_h(X_i-x)(r(X_i)+\epsilon_i^*-r(x))$ and
$B_n=\frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)r(X_i)(\hat{m}_p(x)-m_p(x))-\frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)\frac{m_p(x)}{m_p(X_i)}r(X_i)(\hat{m}_p(X_i)-m_p(X_i))$.

$$\begin{aligned}E(A_n) &= m_p(x)E\left(n^{-1}\sum_{i=1}^{n} K_h(X_i-x)(r(X_i)-r(x))\right)\\ &= m_p(x)\int K_h(X_1-x)\left(r(X_1)-r(x)\right)f(X_1)dX_1\\ &= m_p(x)\int K(\omega)\left(r(x+h\omega)-r(x)\right)f(x+h\omega)d\omega \text{ after a change of variable}\\ &= \frac{h^2}{2}\left(m_p(x)f(x)r''(x)+2m_p(x)f'(x)r'(x)\right)\mu_2(K)+o(h^2). \qquad (20)\end{aligned}$$

Denote B_n^1 and B_n^2 respectively the first and second terms of B_n.

$$\begin{aligned}E(B_n^1) &= E\frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)r(X_i)E_{X_i}\left(\hat{m}_p(x)-m_p(x)\right)\\ &= \frac{1}{2}\mu_2 h_p^2\left(m_p''(x)+2m_p'(x)\frac{g'(x)}{g(x)}\right)E\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i-x)r(X_i)\right)\\ &= r(x)f(x)Bias\left(\hat{m}_p(x)\right)+o(h^2).\end{aligned}$$

Similarly,

$$\begin{aligned}
-E(B_n^2) &= E\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) r(X_i) \frac{m(x)}{m_p(X_i)} E_{X_i}\left(\hat{m}_p(X_i) - m_p(X_i)\right)\right) \\
&= \frac{1}{2}\mu_2 h_p^2 E\left(m_p''(X_i) + 2m_p'(X_i)\frac{g'(X_i)}{g(X_i)}\right) E\left(\frac{1}{n}\sum_{j\neq i} K_h(X_j - x) r(X_j)\frac{m_p(x)}{m_p(X_j)}\right) + o(n^{-1})
\end{aligned}$$

by assumption A1. Hence,

$$\begin{aligned}
-E(B_n^2) &= \frac{1}{2}\mu_2 h_p^2 \left(m_p''(x) + 2m_p'(x)\frac{g'(x)}{g(x)}\right) r(x) f(x) + o(h^2 + n^{-1}) \\
-E(B_n^2) &= r(x) f(x) Bias[\hat{m}_p(x)] + o(h^2 + n^{-1}).
\end{aligned}$$

Since $plim\hat{f}(x) = f(x)$, it follows that $E(\hat{m}(x) - m(x)) \simeq f(x)^{-1} E(A_n + B_n)$. This completes the first part of the proof.

$Var[A_n] = \sigma^2 (nh)^{-1} R(K) f(x) + O(h/n)$. The computation of the variance of B_n and the covariance of A_n and B_n is significantly longer and thus not provided in detail; it is available from the authors. Both $Var[B_n]$ and $Cov(A_n, B_n)$ are found to be the order $O[(Nh_p)^{-1}]$. Again, $Var(\hat{m}(x)) \simeq f(x)^{-2}[Var(A_n) + Var(B_n) + 2Cov(A_n, B_n)]$, which completes the second part of the proof.

2. Under the assumptions A1-A7, $\hat{m}(x)$ has a limiting normal distribution:

$$\sqrt{nh}(\hat{m}(x) - m(x) - B(h)) \to N(0, \Sigma) \tag{21}$$

where $B(h) = \frac{1}{2}\mu_2 h^2 [m_p(x) r''(x) + 2m_p(x) r' \frac{f'(x)}{f(x)}]$ and $\Sigma = \frac{\sigma^2}{f(x)} R(K)$.

Proof. Write $(\hat{m}(x) - m(x))\hat{f}(x) = C_n + D_n + o_p(h_p^2)$ where $C_n = \frac{m_p(x)}{n}\sum_{i=1}^{n} K_h(X_i - x)(r(X_i) - r(x)) + \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x) r(X_i)(\hat{m}_p(x) - m_p(x)) - \frac{1}{n}\sum_{i=1}^{n} K_h(X_i - x)\frac{m_p(x)}{m_p(X_i)} r(X_i)(\hat{m}_p(X_i) - m_p(X_i))$ and $D_n = \frac{m_p(x)}{n}\sum_{i=1}^{n}\frac{\epsilon_i}{m_p(X_i)} K_h(X_i - x)$. From the first part of the proof of the theorem, it can be seen that $E(C_n) = \frac{h^2}{2}\left(m_p(x) f(x) r''(x) + 2m_p(x) f'(x) r'(x)\right)\mu_2(K) + o(h^2)$; somewhat lengthy calculations show that $Var(C_n) = o(h^4) + O(\frac{1}{n_1 h_p + n_2 h_p + ... + n_Q h_p})$. By assumption A7, $n_j h_p \to \infty \ \forall$ $j = 1, ..., Q$; hence the last term of the variance of C_n can be ignored. Combining the expectation and variance of C_n, it follows that

$$\begin{aligned}
C_n &= E(C_n) + o_p(h^2) \\
&= \frac{h^2}{2}\left(m_p(x) f(x) r''(x) + 2m_p(x) f'(x) r'(x)\right)\mu_2(K) + o_p(h^2) \\
&= f(x) B(h) + o_p(h^2);
\end{aligned}$$

Similarly, $E(D_n) = 0$ and $Var(D_n) = (nh)^{-1}\left(\sigma^2 R(K) f(x) + o(1)\right)$. D_n is a triangular array of i.i.d. random variables; thus, under assumption A6, we can apply Liapounov's central limit theorem to obtain $\sqrt{nh}(D_n) \to N(0, f^2(x)\Sigma)$.

Since $plim\hat{f}(x) = f(x)$, it also follows that

$$\sqrt{nh}(\hat{m}(x) - m(x) - B(h)) = \sqrt{nh}\frac{D_n}{\hat{f}(x)} + o_p(1) = \sqrt{nh}\frac{D_n}{f(x)} + o_p(1) \to N(0, \Sigma). \tag{22}$$